Introduction of AI (CS103)- 11 Machine Learning Algorithms 3

Jimmy Liu 刘江

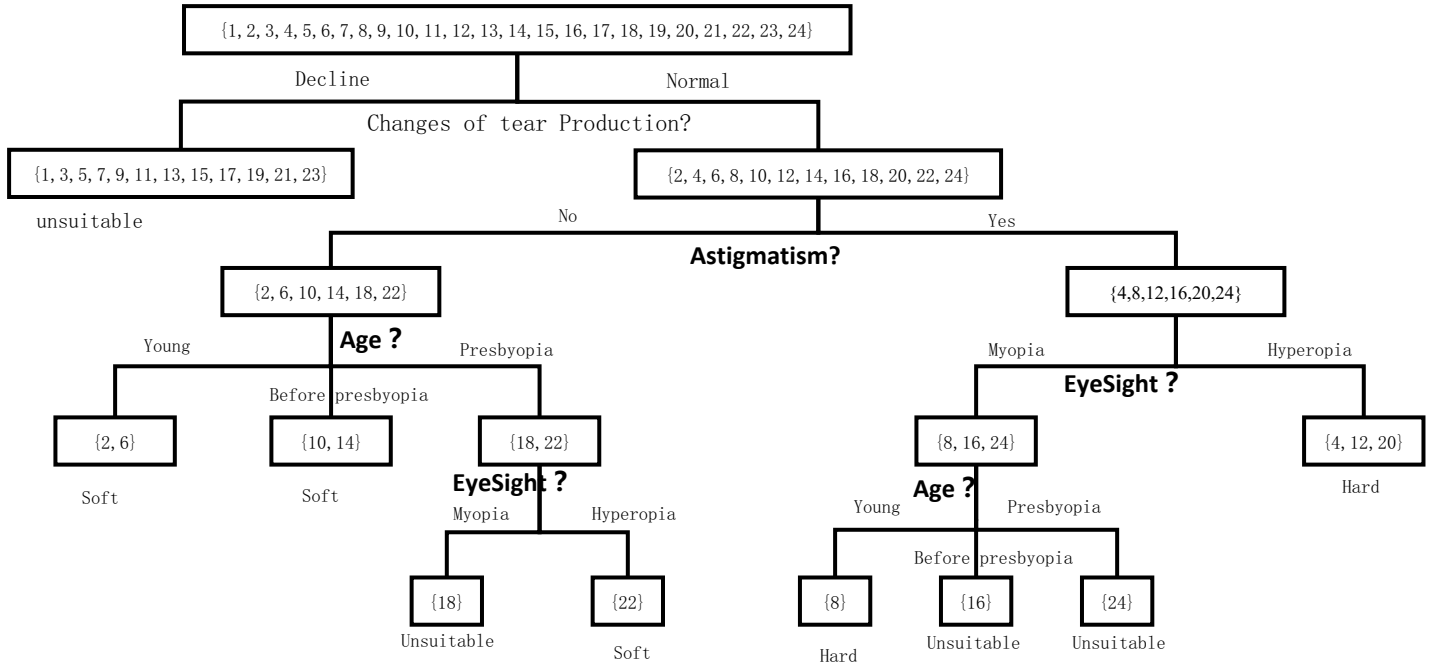
2023-12-01

Lecture 11

1. Reviews of Lecture 9
2. Unsupervised Learning
3. PCA
4. Weakly Supervised Learning

Decision Tree

We can see that decision tree is comprehensible and explainable. We can choose the gain ratio and Gini index according to actual situations.



CS103 HW8

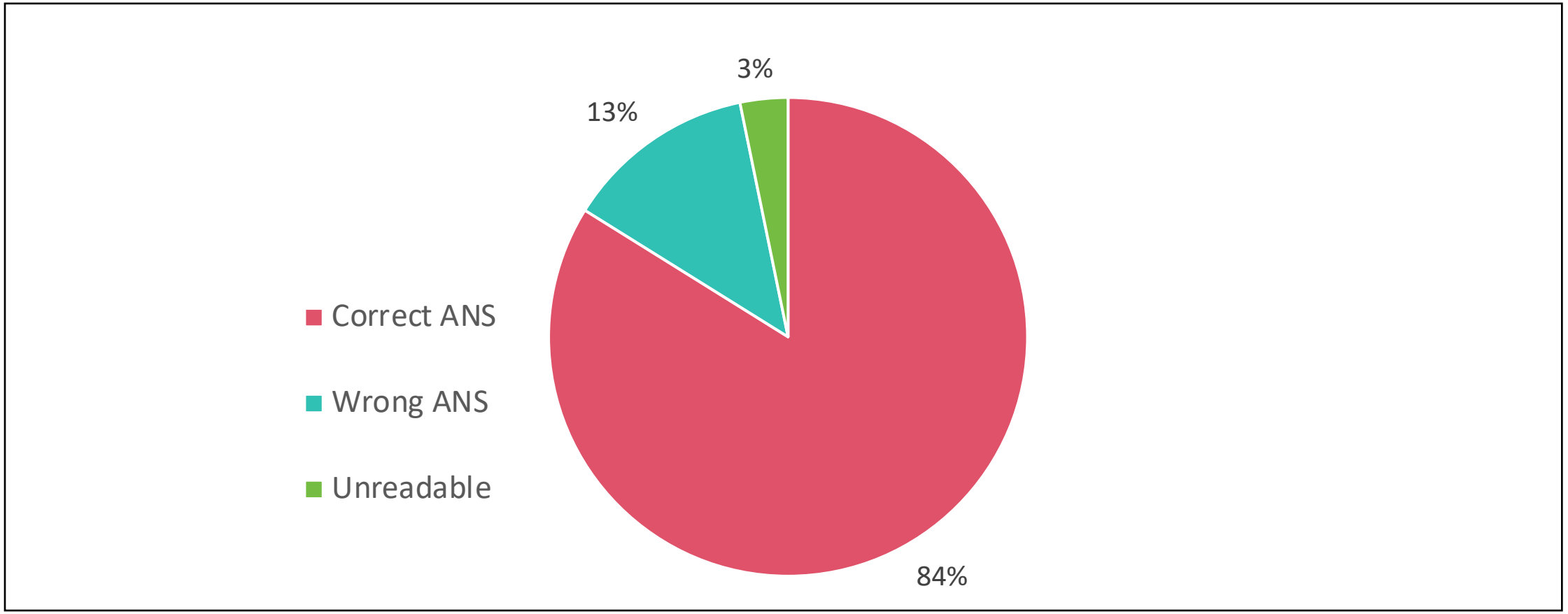
Suppose we now have the following data set, in which flower color and leaf shape are discrete features, and flower types are their labels. Now, we want to classify the flower species in the dataset through decision tree, please answer the following questions.

serial number	flower color	leaf shape	flower type
1	red	circle	Iris-Ame
2	red	strip	Iris-Ame
3	white	acicular	Iris-Ame
4	white	acicular	Iris-Somnus
5	white	strip	Iris-Somnus
6	purple	circle	Iris-Somnus
7	purple	acicular	Iris-XinQ
8	red	circle	Iris- XinQ
9	purple	strip	Iris- XinQ

（1）Please calculate the information gain of flower color and leaf shape as selected features separately.

（2）Please explain which feature should be selected based on the calculation results in (1)?

CS103 HW8



Homework 11-1: How to You Expect AI + Medical Diagnosis to Be When You Graduate ?

Homework 11-2: How do You Suggest CS 103 to Be Taught With “AI +Education” Development Next Year?

Lecture 11

1. Reviews of Lecture 9
2. Unsupervised Learning
3. PCA
4. Weakly Supervised Learning

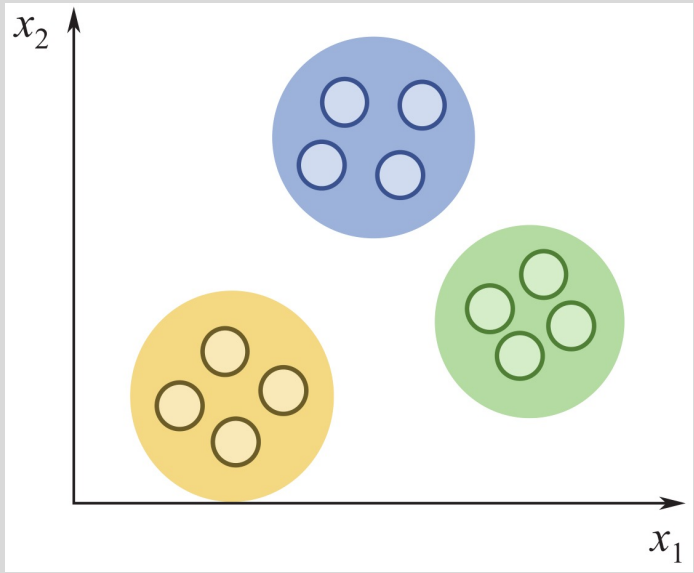
Unsupervised Learning - K-means

K- means

- Simple but most popular partitional algorithm.
- Assume Euclidean space.
- k clusters: $C_1, C_2, \ldots, C_k$
- Minimise the sum of squared distances to the centroid of clusters over all k clusters:

$$E = \sum_{i=1}^{k} \sum_{x \in C_i} ||x - \mu_i||_2^2$$

which represents the tightness of the samples in the cluster around the cluster mean vector μ, and the smaller the square error, the higher the similarity of the samples in the cluster.



Clustering

K-means

- **Main Procedure :**

  ① Determine the number of clusters k (plan to divide the data into k classes);

  ② Randomly determine k initial points as the center of mass (randomly selected within the range of data boundaries);

  ③ Calculate the distance to k centroids for each data instance, select the centroid of the smallest distance, and assign it to the cluster corresponding to the centroid until all the data in the data set are allocated to k clusters, and update the centroids of k clusters to the average of all points in the cluster;

  ④ Repeat step ③ to reassign each data instance to a new center of mass until a termination condition is reached, such as no further changes in the results of all data allocations.

K-means

Example of K-means algorithm application

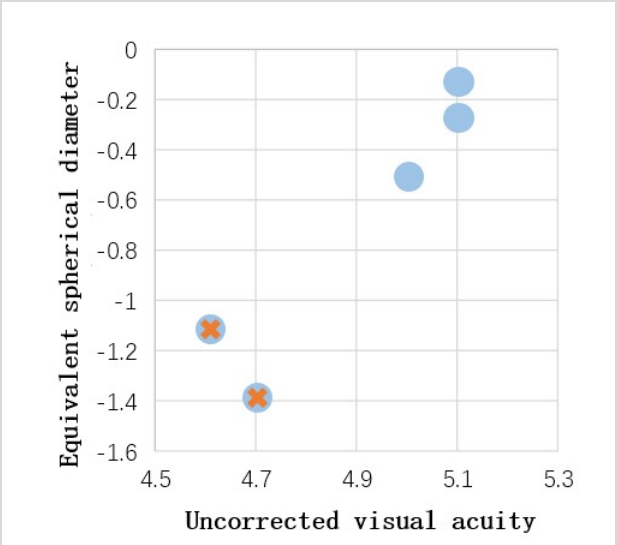
The examination of myopia mainly includes the examination of naked eye vision and equivalent spherical lens. The naked eye visual acuity examination mainly checks the degree of myopia in the naked eye. The equivalent spherical lens is a spherical lens that can convert astigmatism into a similar optical effect, indicating the refractive state of the eye. The following is a cluster analysis of myopia and non-myopia based on naked eye visual acuity and equivalent spherical lens.

Sample number	Uncorrected visual acuity	Equivalent spherical diameter
1	4.7	-1.38
2	4.6	-1.13
3	5	-0.5
4	5.1	-0.25
5	5.1	-0.13

K-means

Example of K-means algorithm application

Step 1: Assume the number of clusters k=2, and randomly select samples 1 and 2 as the initial clustering centers, that is, the initial mean vector $\mu_1 = (4.7, -1.38), \mu_2 = (4.6, -1.13)$. The initial situation is shown in the figure on the right, with the blue circle representing the sample points and the orange cross representing the cluster center.



K-means

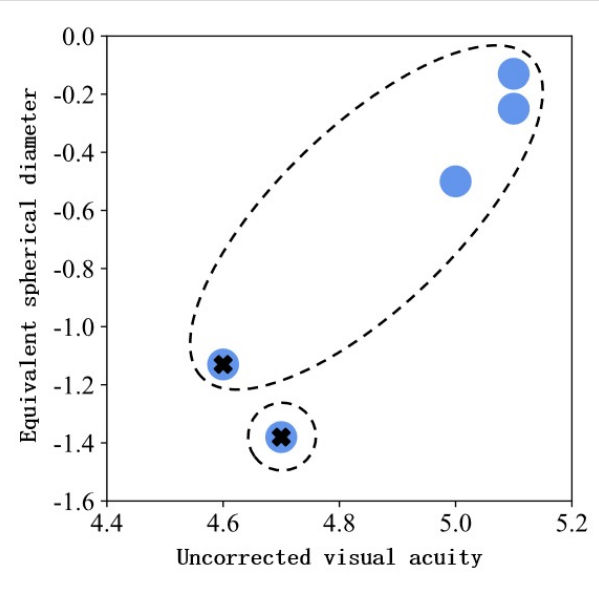
Example of K-means algorithm application

Step 2: Calculate the distance from each sample point to the initial cluster center separately. And regroup according to distance, the results are as follows:

Sample points contained in cluster center A: Sample 1.

Sample points contained in cluster center B: sample 2 ~ sample 5.

Sample number	Cluster center A (4.7, -1.38)	Cluster center B (4.6, -1.13)
1	0	0.2693
2	0.2693	0
3	0.9297	0.7463
4	1.1987	1.0121
5	1.3124	1.1180



K-means

Example of K-means algorithm application

Step 3: according to the result of regrouping computing new clustering center, A clustering center of the mean vector is: $(4.7, -1.38)$, the clustering center B new mean vector $\left(\frac{4.6+5+5.1+5.1}{4}, \frac{-1.13-0.5-0.25-0.13}{4}\right) = (4.95, -0.5025)$. Then recalculate the distance.

Sample number	Cluster center A (4.7, -1.38)	Cluster center B (4.95, -0.5025)
1	0	0.9124
2	0.2693	0.7185
3	0.9297	0.0501
4	1.1987	0.2937
5	1.3124	0.4016

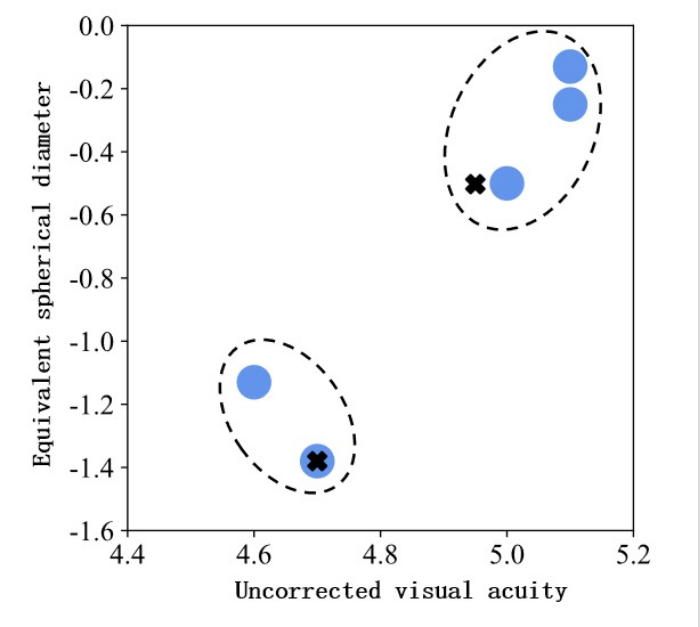
K-means

Example of K-means algorithm application

Step 3: Regroup according to distance, and the result is as follows:

Cluster center A contains sample points: sample 1, sample 2.

Cluster center B contains sample points: sample 3, sample 4, sample 5.



K-means

Example of K-means algorithm application

Step 4: to recalculate the center of mass, the clustering center is A new mean vector is $\left(\frac{4.7+4.6}{2}, \frac{-1.38-1.13}{2}\right) = (4.65, -1.255)$.

New clustering center B mean vector is: $\left(\frac{5+5.1+5.1}{3}, \frac{-0.5-0.25-0.13}{3}\right) = (5.0667, -0.2933)$.

Then recalculate the distance as follows:

Sample number	Cluster center A (4.65, -1.255)	Cluster center B (5.0667, -0.2933)
1	0.1346	1.1469
2	0.1346	0.9581
3	0.8322	0.2172
4	1.1011	0.0546
5	1.2117	0.1667

K-means

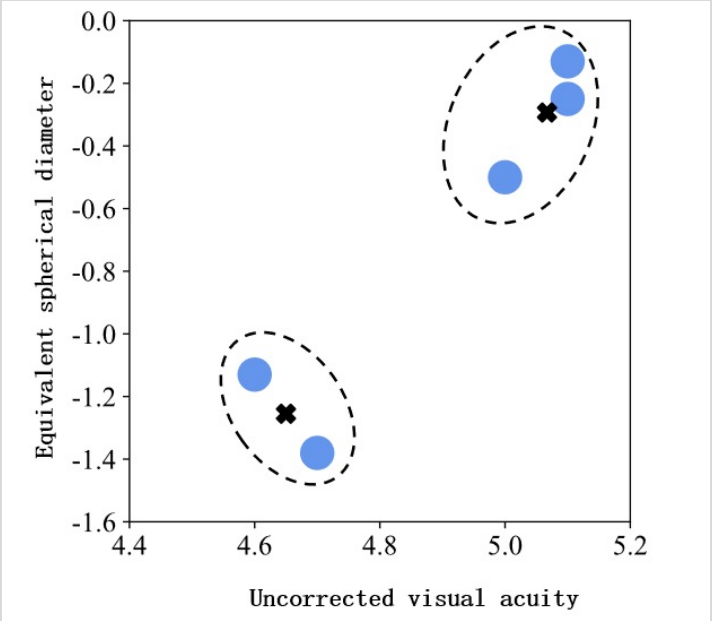
Example of K-means algorithm application

Step 4: Regroup according to distance, the result is as follows:

Cluster center A contains sample points: sample 1, sample 2.

Cluster center B contains sample points: sample 3, sample 4, sample 5.

From then on, the clustering result does not change and the clustering ends. Therefore, the clustering results are shown in the figure on the right. Samples 1 and 2 belong to the same category, and samples 3 to 5 belong to the same category.



Q1: How do K-means deal with empty clusters

In K-means clustering algorithm, empty cluster refers to the case that no sample points are divided into the cluster after a certain round of iteration.

Lecture 11

1. Reviews of Lecture 9
2. Unsupervised Learning
3. PCA
4. Weakly Supervised Learning

Principle Component Analysis

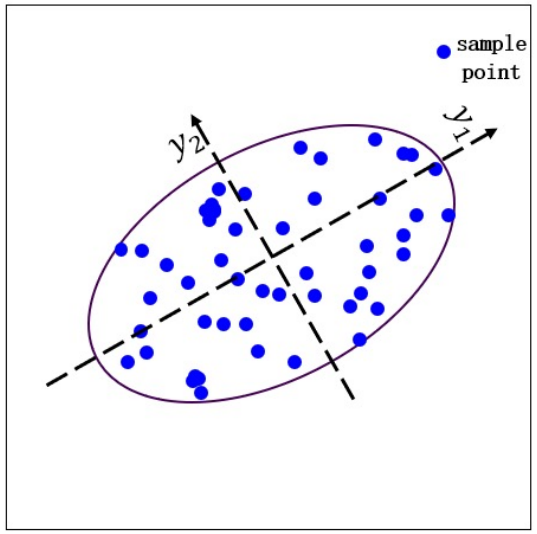
- Apply linear transformation to features to make the new features linearly independent :
  - the dependency among features can be expressed by a covariance matrix: the element in the i-th row and j-th column is the correlation between the i-th and j-th feature.
  - therefore, if off-diagonal elements are zero, then all features are linearly independent.

- Suppose we have training examples represented by a normalized N × d matrix X. We aim to make the covariance matrix of XX^T after transformation is a diagonal matrix.

- For X, eigenvectors of the associated covariance matrix meet the above requirements.

Principle Component Analysis

The objective of PCA is to find a new set of orthogonal bases $\{u_1, u_2, \ldots, u_k\}$, so that after the data points are projected on the plane formed by the orthogonal basis, the distance between the data is the largest, that is, the variance between the data is the largest. If the variance of the data is the largest after projection on each orthogonal basis, it is also satisfied that the projection distance is the largest on the plane formed by the orthogonal basis.



Example of principal component analysis

Principle Component Analysis

- **Main Procedure** :

Input: Sample set $D = \{x_1, x_2, \cdots, x_{\mathrm{m}}\}$;

The dimension of a low-dimensional space d'。

Procedure :

1. Centralize all samples : $x_{\mathrm{i}} \leftarrow x_i - \frac{1}{m} = \sum_{i=1}^{m} x_i$;
2. Calculate the covariance matrix $\mathbf{X}\mathbf{X}^{\mathrm{T}}$ of the samples ;
3. The eigenvalue decomposition of covariance matrix $\mathbf{X}\mathbf{X}^{\mathrm{T}}$ is obtained.;
4. Take the eigenvector $\boldsymbol{w}_1, \boldsymbol{w}_2, \cdots, \boldsymbol{w}_{d'}$ corresponding to the largest d' eigenvalues.

Output: Projection matrix $\boldsymbol{W} = (\boldsymbol{w}_1, \boldsymbol{w}_2, \cdots, \boldsymbol{w}_{d'})$

Principle Component Analysis

Example of PCA application

We take the prediction of myopia severity as an example. When patients with myopia go to the hospital for treatment, the doctor will collect the patient's age, naked eye, corrected vision, equivalent spherical lens, spherical lens, astigmatism and other information. We select 10 information that will affect vision as indicators, as shown in the figure, and form a 16×10 matrix from 16 patients as the original input of PCA.

Principle Component Analysis

Example

Sample number	Age	Naked vision	Corrected vision	Correction method	Equivalent spherical mirror	Spherical mirror	Astigmatism	Steep meridian corneal diopter (D)	Flat meridian diopter (D)	Axial AL	nearsightedness
1	6	5.1	0	0	0.13	0.25	-0.25	42.03	42.78	22.2	2
2	6	4.9	0	0	0.13	0.25	-0.25	42.13	43.44	22.2	2
3	6	5	0	0	0.25	0.25	0	41.31	42.51	22.86	2
4	6	5	0	0	0.25	0.25	0	43.55	43.83	22.58	2
5	6	5	0	0	0.25	0.5	-0.5	40.47	41.31	23.9	2
6	6	5	0	0	0.25	0.25	0	43.77	44.82	22.19	2
7	7	5	0	0	0.5	0.5	0	43.6	45.06	22.84	2
8	7	5.1	0	0	0.63	0.75	-0.25	44.23	45.55	21.65	2
9	7	5	0	0	0.63	1	-0.75	41.62	42.78	23.13	2
10	7	5	0	0	0.63	0.75	-0.25	43.38	45.36	21.81	2
11	7	5.1	0	0	0.75	0.75	0	43.49	44.58	22.59	2
12	6	4.3	4.9	1	-2.5	-1.5	-2	41.93	44.06	24.11	1
13	6	4.5	4.8	1	-2	-1.75	-0.5	43.21	44.23	23.32	1
14	6	4.4	0	0	-1.25	-1	-0.5	43.32	44.06	23.28	1
15	6	4.7	4.9	1	-1.13	0.5	-3.25	42.24	45.49	23.01	1
16	6	4.8	0	0	-1.13	0	-2.25	39.47	42.03	23.94	1

Principle Component Analysis

Example of PCA application

Step 1 : centralize all features. Find the mean $\mu_j = \frac{1}{m}\sum_{j=1}^{m} x_j$. For example, the mean of feature 1 (age) μ1 = 6.3125, the mean of feature 2 (naked eye vision) μ2 = 4.86875, and the means of other features are calculated accordingly. Then, for all the samples, each feature is subtracted from its own mean to get the de-averaged feature.

Step 2 : Calculate the covariance matrix. $\sum = \frac{1}{m-1}\begin{bmatrix} \mathrm{cov}(\boldsymbol{X}_1, \boldsymbol{X}_1) & \mathrm{cov}(\boldsymbol{X}_1, \boldsymbol{X}_2) & \dots & \mathrm{cov}(\boldsymbol{X}_1, \boldsymbol{X}_n) \\ \mathrm{cov}(\boldsymbol{X}_2, \boldsymbol{X}_1) & \mathrm{cov}(\boldsymbol{X}_2, \boldsymbol{X}_2) & \dots & \mathrm{cov}(\boldsymbol{X}_2, \boldsymbol{X}_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}(\boldsymbol{X}_n, \boldsymbol{X}_1) & \mathrm{cov}(\boldsymbol{X}_n, \boldsymbol{X}_2) & \dots & \mathrm{cov}(\boldsymbol{X}_n, \boldsymbol{X}_n) \end{bmatrix}$

Principle Component Analysis

Example of PCA application

Step 3: Calculate the eigenvalues of the covariance matrix and their corresponding eigenvectors.

Step 4: The eigenvalues of the covariance matrix are calculated and the eigenvectors corresponding to the first k eigenvalues are extracted to form a new matrix.

Sample number	Eigenvalue	Contribution rate	Cumulative contribution rate
1	5.391263	53.91263	53.91263
2	2.436536	24.36536	78.27799
3	1.085323	10.85323	89.13122
4	0.565423	5.654227	94.78545
5	0.323205	3.232055	98.01751
6	0.130434	1.30434	99.32185
7	0.04464	0.446401	99.76825
8	0.023149	0.23149	99.99974
9	2.59E-05	0.000259	100
10	4.31E-07	4.31E-06	1.00E+02

Principle Component Analysis

Example of PCA application

PCA reduces the data from n dimension to k dimension, how to choose the appropriate k?

The general selection criterion is: the ratio of variance before and after the projection, which is used as the selection criterion of k value. Specifically, we expect: $\frac{Var_{X_{project}}}{\mathrm{Var}_X} \geq q$, and q is generally 0.99. Therefore, the number of principal components k is the minimum k that satisfies the above conditions $\frac{Var_{X_{project}}}{\mathrm{Var}_X} = \frac{\sum_{j=1}^{k} \lambda_j}{\sum_{j=1}^{n} \lambda_j} \geq 0.99.$

Sample number	Eigenvalue	Contribution rate	Cumulative contribution rate
1	5.391263	53.91263	53.91263
2	2.436536	24.36536	78.27799
3	1.085323	10.85323	89.13122
4	0.565423	5.654227	94.78545
5	0.323205	3.232055	98.01751
6	0.130434	1.30434	99.32185
7	0.04464	0.446401	99.76825
8	0.023149	0.23149	99.99974
9	2.59E-05	0.000259	100
10	4.31E-07	4.31E-06	1.00E+02

Principle Component Analysis

Example of PCA application

In this application example, take the first 6 principal component eigenvectors. Large eigenvalues correspond to large proportions of components, such as equivalent spherical mirror, flat meridian diopter, and spherical mirror in the first, second and third principal component eigenvectors respectively.

Primary index	First principal component eigenvector	Second principal component eigenvector	Third principal component eigenvector	Fourth principal component eigenvector	Fifth principal component eigenvector	Sixth principal component eigenvector
Age	0.26281	0.237389	0.357388	0.762276	0.131525	-0.36259
Naked vision	0.388745	-0.11119	0.19598	-0.26327	0.385393	0.07327
Corrected vision	-0.3638	0.251086	0.214078	-0.13166	0.474003	-0.00297
Correction method	-0.3637	0.251634	0.20892	-0.12972	0.481965	-0.00997
Equivalent spherical mirror	0.420219	-0.05419	0.108914	-0.0549	0.190441	0.239719
Spherical mirror	0.351448	-0.09759	0.504501	-0.16399	-0.04209	0.304118
Astigmatism	0.302468	0.047975	-0.61487	0.158214	0.473874	0.008925
Steep meridian corneal diopter	0.144881	0.5619	-0.2743	0.019979	-0.03309	0.450028
Flat meridian diopter	0.041456	0.613562	0.144559	-0.03769	-0.32609	0.169639
Axial AL	-0.31762	-0.313	0.061686	0.508629	0.08471	0.69406

Q2: Why does PCA use the eigenvector matrix of the covariance matrix as the projection matrix?

Homework Week 11-3

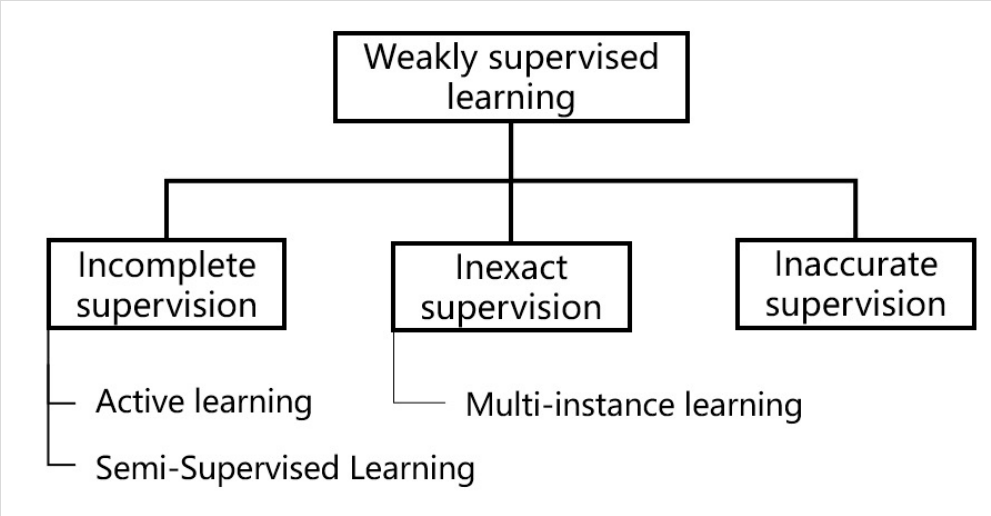
- Suppose there are the following 8 points: (5,1),(5,2),(4,1),(4,2),(1,3),(1,4),(2,3),(2,4). Cluster them using K-means algorithm. Let the initial clustering centers be (0,4) and (3,3). Please write down the detailed calculation process.

Lecture 11

1. Reviews of Lecture 9
2. Unsupervised Learning
3. PCA
4. Weakly Supervised Learning

Weakly Supervised Learning

- Weakly supervised learning is a branch of machine learning where the data used to train a model is limited, noisy, or labeled inaccurately compared to traditional supervised learning. Due to **the lack of effective supervision information**, it is expected that machine learning techniques can actively learn effective data feature representations under weak supervision.
- Classification according to the degree of data labeling: incomplete supervision, inexact supervision, inaccurate supervision.

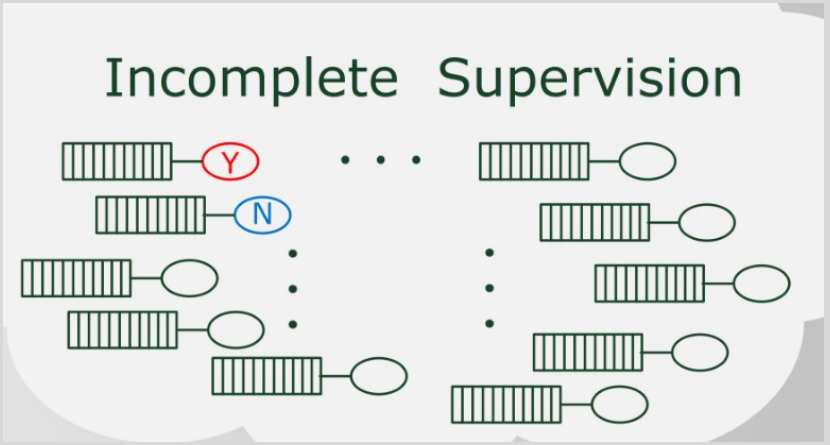


Types of weakly supervised learning

Weakly Supervised Learning

Weakly supervised learning types

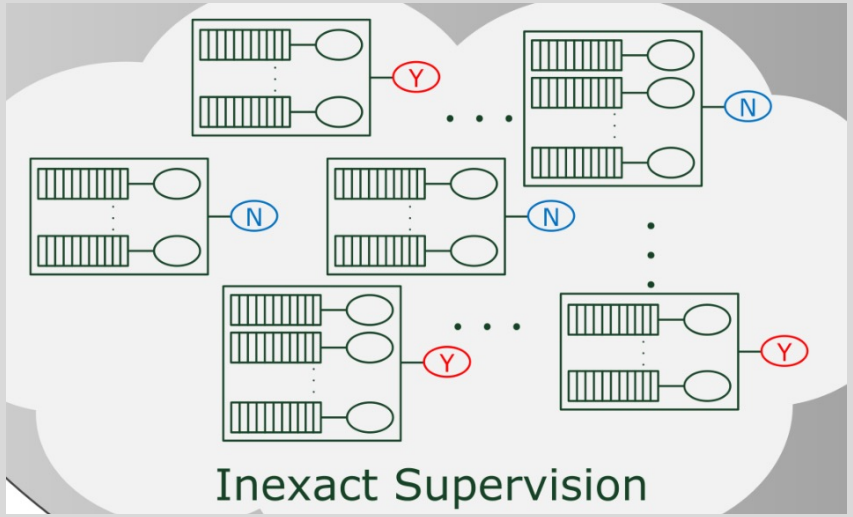
- **Incomplete supervision:** Only a subset of the total training data has supervision labels. For example, in the task of disease recognition and screening based on medical images, the number of medical images with accurate labeling is very small, and most of the medical image data is not labeled.



Weakly Supervised Learning

Weakly supervised learning types

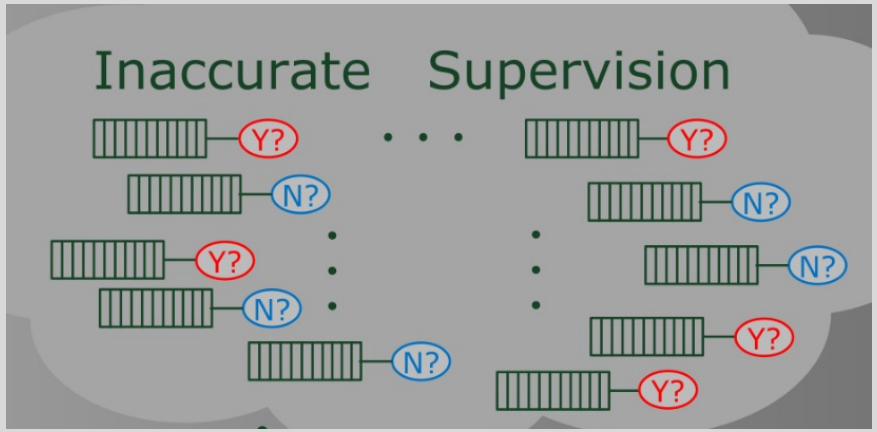
- **Inexact supervision:** Only coarse-grained labels are given for the data. Taking the task of disease recognition and screening based on medical images as an example, doctors usually only give whether there are abnormal areas in a single medical image, that is, image-level annotation, rather than marking pathological areas in medical images, that is, pixel-level annotation or object-level annotation.



Weakly Supervised Learning

Weakly supervised learning types

- **Inaccurate supervision:** The problem is that a given label is not always true. Take the task of disease recognition and screening based on medical images as an example, in which the quality of manual medical image labeling largely depends on the knowledge, work intensity, experience and clinical training of the doctor who read the film, which may make the given label **not always true**, especially for some medical images that are difficult to be accurately classified.

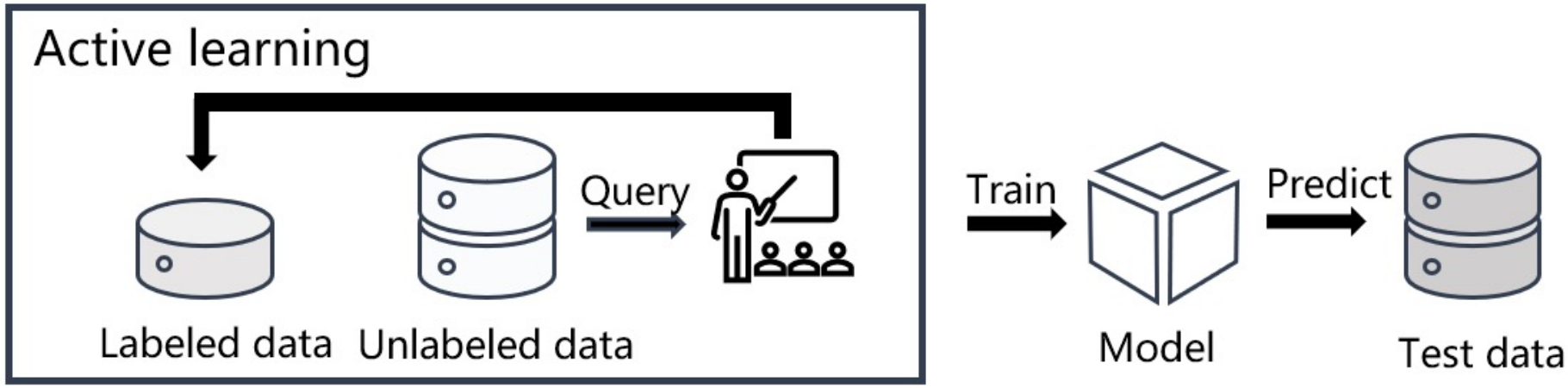


Q3: Which is an accurate description of the "weakly supervised learning" paradigm in machine learning?

A. Receive a set of labeled data and make predictions about all the unknown data.
B. It is designed to use a limited number of training samples to learn some information.
C. Be able to solve a task even if no training samples of the task are obtained.
D. Belongs to the class of supervised learning algorithms, in addition to using unlabeled data for training.

Example 1: Active Learning

- Active learning assumes that real labels can be **queried** from unlabeled instances.
- **Scenario:** Given a small amount of labeled data and a large amount of unlabeled data, the general idea of active learning is to obtain the "difficult" classified sample data through machine learning, label it manually, and then train the manually labeled data again to improve the model performance.
- **Goal:** The goal is to minimize tag costs, that is, to minimize the number of queries.



Active learning

Example 2: Semi-Supervised Learning

- In the case of a small amount of labeled data and a large amount of unlabeled data, part of the training is labeled and part is unlabeled.
- Basic assumptions of data distribution:
  - Smoothness assumption
  - Cluster assumption
  - Manifold assumption

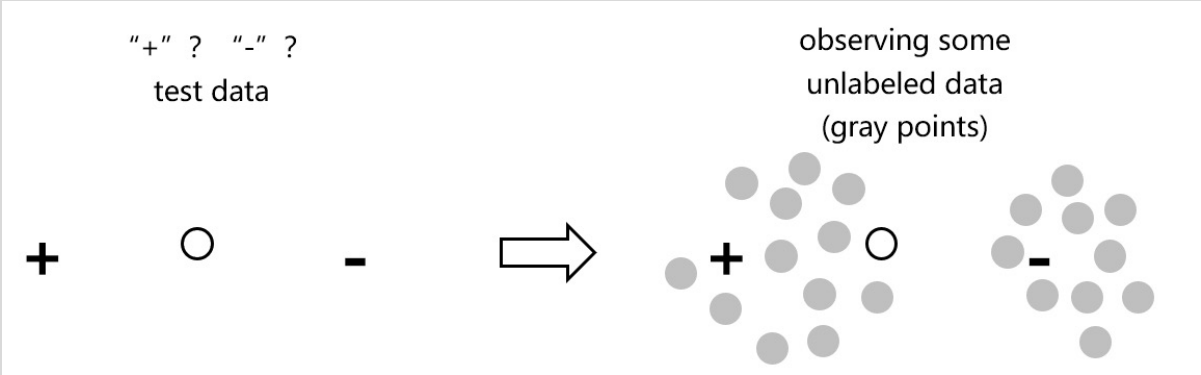
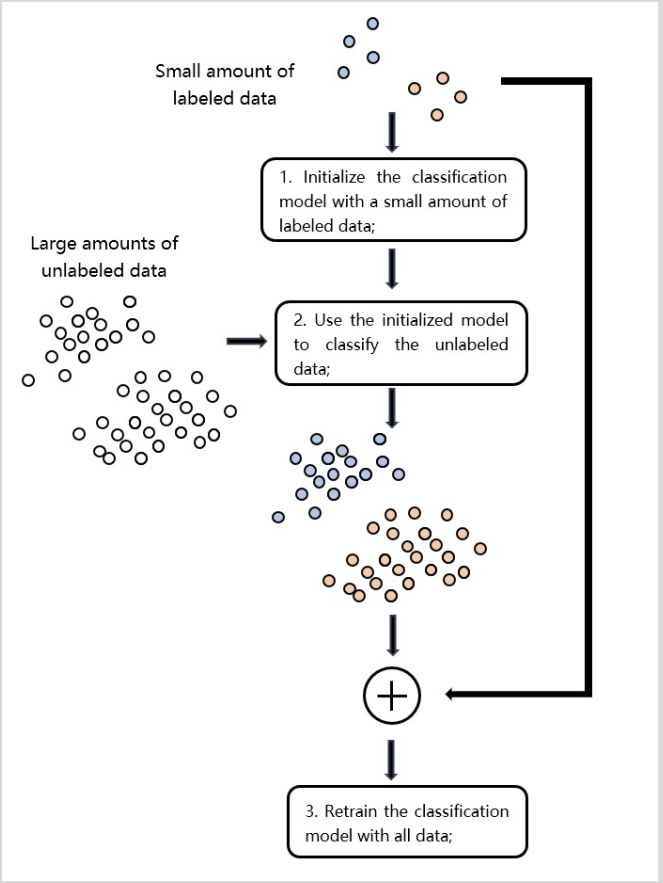


Illustration of the usefulness of unlabeled data

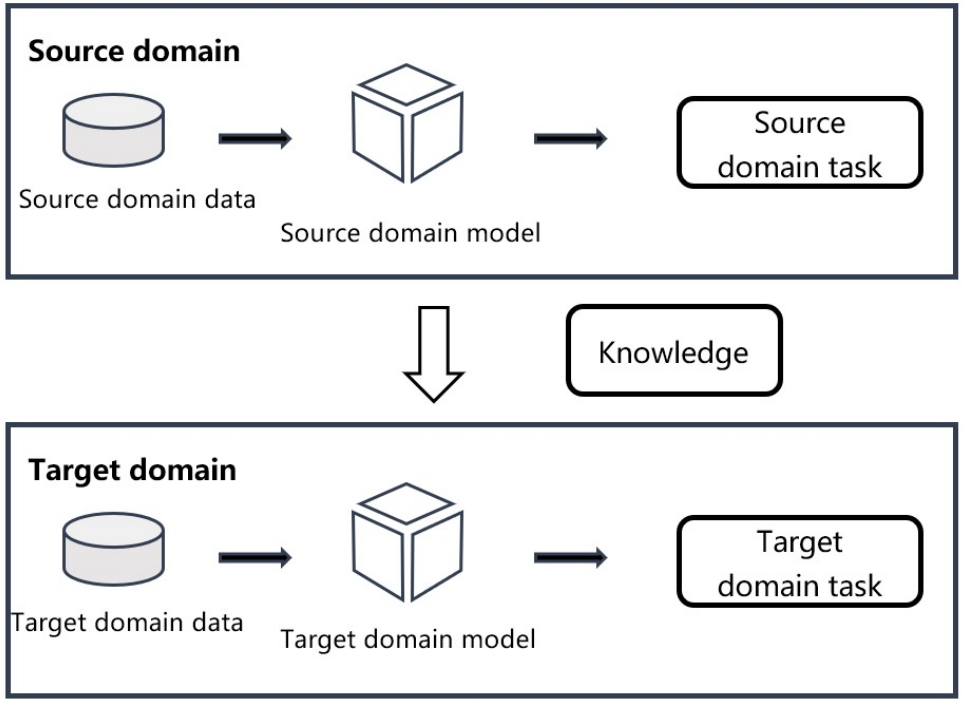
Semi-Supervised Learning

- **Process:** (Taking binary task as an example)
  - Train the model with a small amount of labeled data;
  - The initialized model is used to label the unlabeled data and obtain false labels.
  - The model is trained with all the data, both labeled and pseudo-labeled.



Example 3: Transfer Learning

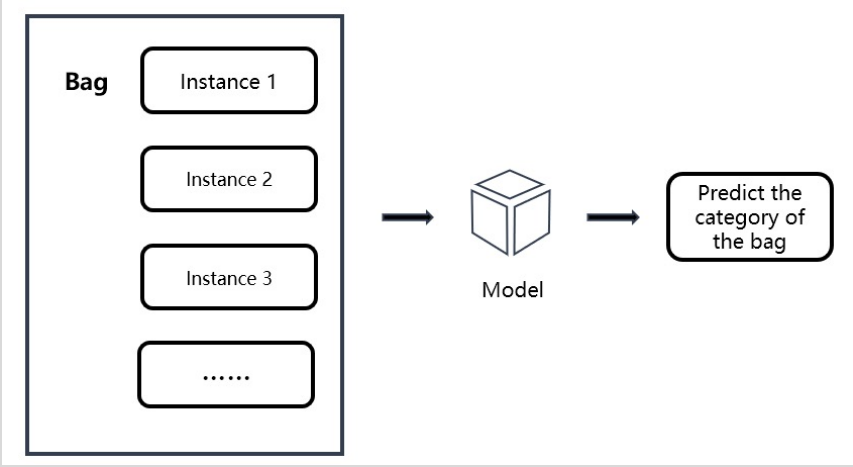
- **Scenario:** The training data of the target task is small, while a related task has a large amount of training data. Although the distribution of the two types of training data is different, the target task is also different, but some generalization knowledge can be learned in the related task, which will be helpful to the target task.
- **Goal:** Transfer generalizable knowledge of the related task to the target task.



Transfer learning

Example 4: Multi-instance learning

- **Scenario:** Taking binary classification task as an example, assuming that all examples in a bag have at least one positive example, the bag is marked as positive; Conversely, if all examples in a bag are negative, the bag is marked as negative, and the purpose of learning is to predict the category of the new package.
- **Application scenarios:** image classification, text classification, spam detection, medical diagnosis, face detection, object tracking, etc.



Multi-instance learning

Q4:Which of the following statements about weakly supervised learning is not true?

A. Weakly supervised learning only introduces labeling knowledge to part of the samples.
B. Weakly supervised learning is equivalent to semi-supervised learning.
C. The core idea of transfer learning is to use the experience gained on task A to solve similar task B.
D. Semi-supervised learning gradually expands unlabeled data by learning labeled data.

Introduction of AI (CS103)- 11 Machine Learning Algorithms 3

Jimmy Liu 刘江

2023-12-01